International Journal of Innovative Trends in Engineering Science and Technology
A Peer-Reviewed International Research Journal (IJITEST)
www.ijitest.org | E-ISSN: XXXX-XXXX

Deep Reinforcement Learning for Dynamic Resource Management in Ephemeral Edge Computing Networks

[1]Dr. Ch. Swapna Priya, [2]Dr. S. NagaMallik Raj, [3]Mahamed Mastan Jani, [4]Bharath Karthik Mycherla, [5]Surya Teja Medisetty, [6]Kalpana Pulipati

[1,2]Associate Professor, [3,4,5,6]Student, B.Tech (Final Year)
[1-6]Department of Computer Science and Engineering,
Vignan's Institute of Information Technology (A), Duvvada, Visakhapatnam, AP, INDIA

Published online: 25 April 2026

Abstract – Efficient resource orchestration in modern edge computing deployments is increasingly challenged by node mobility, stochastic workloads, and limited energy budgets. Conventional static and heuristic scheduling methods are fundamentally inadequate for volatile environments such as UAV swarms and vehicular ad hoc networks, where topology and resource availability evolve continuously. This paper proposes a novel adaptive resource management framework grounded in Proximal Policy Optimization (PPO), a state-of-the-art Deep Reinforcement Learning (DRL) algorithm, tailored for ephemeral edge computing scenarios. The resource allocation problem is rigorously formalized as a Markov Decision Process (MDP) that jointly accounts for end-to-end task latency, cumulative energy expenditure, load distribution fairness, and Service Level Agreement (SLA) compliance. Through iterative interaction with a realistic simulation environment encompassing 20 mobile UAV nodes, the PPO agent acquires nuanced allocation policies that balance competing performance objectives. Our key novelty lies in a composite reward signal that explicitly penalizes battery depletion events, discouraging greedy local processing in favor of energy-balanced, network-lifetime-aware decisions. Experimental results demonstrate that the proposed PPO-based framework reduces SLA violations by approximately 30% and extends network operational lifetime by up to 47% compared to Deep Q-Network (DQN) baselines and classical static schedulers.

Index Terms – Deep Reinforcement Learning, Proximal Policy Optimization (PPO), Edge Computing, Dynamic Resource Allocation, Markov Decision Process, UAV Networks, Energy Efficiency, SLA Compliance.

I. INTRODUCTION

The proliferation of Internet of Things (IoT) devices—projected to exceed 29 billion globally by 2030—has fundamentally transformed contemporary computing paradigms [1]. Latency-sensitive applications including autonomous vehicular control, industrial automation, augmented reality, and real-time health monitoring impose stringent end-to-end delay requirements that conventional cloud architectures are ill-equipped to satisfy. Edge computing has consequently emerged as a critical enabler, shifting computation closer to data sources to reduce propagation delays and alleviate backbone network congestion [2].

Despite these advantages, edge computing environments introduce formidable resource management challenges. Unlike the relatively stable infrastructure of cloud data centers, edge deployments are inherently dynamic: nodes exhibit physical mobility, available computational capacity fluctuates with workload patterns, and energy reserves are strictly finite. These difficulties are accentuated in ephemeral deployments—such as unmanned aerial vehicle (UAV) swarms and vehicular networks—where participating nodes continuously join and leave the network.

Traditional resource management strategies, encompassing heuristic schedulers and static allocation policies, are fundamentally reactive and fail to accommodate real-time environmental shifts. Reinforcement Learning (RL) offers a principled alternative by enabling autonomous agents to optimize decision-making through sustained environmental interaction. Deep Reinforcement Learning (DRL) extends this capability by integrating deep neural networks, making the approach tractable for high-dimensional state spaces characteristic of real-world edge scenarios [3].

Among DRL algorithms, Proximal Policy Optimization (PPO) has demonstrated superior stability and sample efficiency by constraining policy update magnitudes through a clipped surrogate objective. Motivated by these properties, this paper proposes a PPO-based DRL framework for adaptive resource management in ephemeral edge computing environments. Our primary contributions are: (1) A formal MDP model capturing the full resource allocation problem in UAV-based edge networks, incorporating latency, energy, load balance, and SLA objectives; (2) A novel composite reward function that explicitly penalizes battery depletion events; (3) A rigorous comparative evaluation against Static Allocation, Local Greedy, Cloud Greedy, and DQN baselines; (4) Detailed hyperparameter documentation enabling full experimental reproducibility.

II. RELATED WORK

DRL has attracted substantial research attention as a solution to dynamic resource allocation problems across cloud and edge computing domains. A progression of investigations has examined RL-driven approaches for optimizing computational resource utilization, latency, and energy efficiency.

A. Cloud Resource Management

Early efforts applied RL to virtual machine placement and workload scheduling in cloud data centers, achieving measurable improvements in resource utilization [2]. These approaches, however, presupposed stable infrastructure and unconstrained power—conditions absent in mobile edge environments. Parallel work in wireless communications employed RL for spectrum allocation and traffic management, demonstrating latency improvements without addressing node mobility or energy constraints.

B. Edge Task Offloading

Recent investigations have examined task offloading strategies between IoT devices and edge servers [4]. Although offloading can reduce processing delay, these studies typically assumed stationary, perpetually-powered edge nodes. DQN-based frameworks showed promise but exhibited instability in highly dynamic deployments, attributed to the sensitivity of Q-value estimation to non-stationary target distributions [3].

C. Policy Gradient Methods in Edge Computing

PPO-based policy optimization methods address DQN instability through a clipped surrogate objective that prevents large, potentially destabilizing policy updates [3]. Wang et al. [5] demonstrated that PPO achieves superior convergence behavior compared to vanilla policy gradient methods in network resource management tasks. However, existing PPO-based edge computing studies have not simultaneously addressed node mobility, battery depletion, and dynamic topology variation—the precise gap this paper targets.

III. PROPOSED METHOD

The proposed system comprises a collection of mobile UAV edge nodes capable of executing computational tasks submitted by ground-level IoT devices. A centralized DRL agent monitors global network state and issues resource allocation decisions in real time. The architecture encompasses three logical planes: an IoT perception plane generating task workloads, a mobile edge computing plane executing tasks across UAV nodes, and a cloud offloading plane serving as a fallback for compute-intensive tasks.

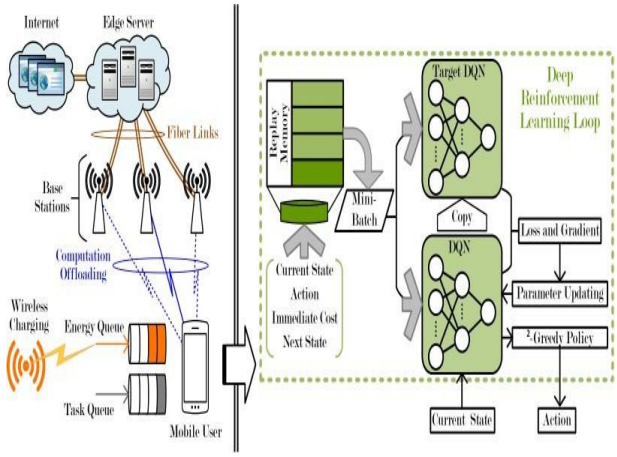

Fig. 1. System architecture of the proposed DRL-based dynamic resource management framework.

A. Markov Decision Process Formulation

The resource allocation problem is cast as a finite-horizon MDP. State Space: st = {B(t), Q(t), τk, Lc(t)}, capturing per-node battery levels B(t), task queue occupancy Q(t), current task attributes τk, and link capacity estimates Lc(t). Action Space: at □ {0, 1, …, N, C}, denoting local processing (0), offloading to one of N peer nodes, or cloud forwarding (C). Reward Function: rt = −α·Lk − β·Etotal − ω·Idrop − η·Ψ(B), where Ψ(B) penalizes battery depletion below threshold βmin.

B. PPO Actor-Critic Architecture

PPO overcomes DQN instability through a clipped surrogate objective: L^CLIP(θ) = Et[min(pt(θ)Ât, clip(pt(θ), 1−ε, 1+ε)Ât)], where pt(θ) is the probability ratio between new and old policies, Ât is the advantage estimate, and ε is the clipping parameter (typically 0.1–0.2). The framework employs an actor network πθ(a|s) and critic network Vφ(s), both implemented as fully-connected networks with [256, 128, 64] hidden units and ReLU activations.

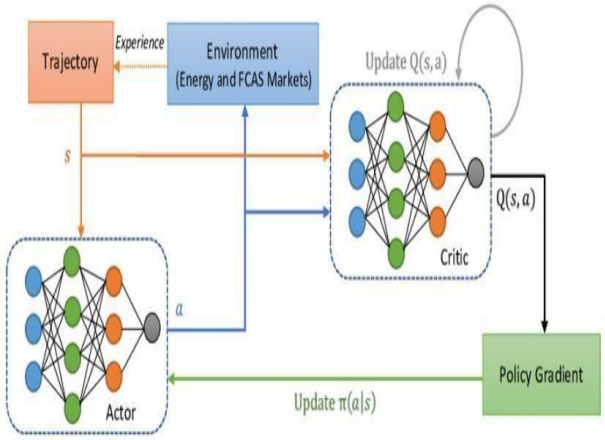

Fig. 2. Reinforcement learning workflow: agent-environment interaction for PPO-based resource allocation.

C. Network and Energy Models

The network comprises N = {1, 2, …, N} mobile UAV nodes. Each node n has processing capacity F_n (CPU cycles/s) and time-varying battery level $B_n(t)$. Local processing latency: Llocal = (Dk × Ck) / F_n. Offload latency: Loffload = Dk /

R_n,m + (Dk × Ck) / Fm. Local energy: Elocal = κ × F_n^2 × (Dk × Ck). Transmission energy: Eoffload = Ptx × (Dk / R_n,m). Total energy: Etotal = Elocal + Eoffload.

IV. RESULTS AND DISCUSSIONS

Experiments were conducted in a simulated 1 km × 1 km area containing 20 UAV nodes following a Random Waypoint Mobility (RWM) model with maximum speed 15 m/s. Each node operated at 2.4 GHz with a 5,000 mAh battery. IoT tasks followed a Poisson arrival process (8 tasks/s) with workloads from [0.1, 1.0] GHz·s and 150 ms SLA deadlines. The PPO agent was compared against Static Allocation (round-robin), Local Greedy, Cloud Greedy, and DQN baselines.

A. Hyperparameter Configuration

PPO agent settings: Actor/Critic Networks FC [256, 128, 64] ReLU; Learning Rate = 3×10^{-4} (Adam); Discount Factor γ = 0.99; PPO Clipping ε = 0.20; GAE λ = 0.95; Entropy Coefficient = 0.01; Mini-Batch Size = 64; PPO Epochs per Update = 10; Reward Weights (α, β, ω, η) = (1.0, 0.5, 5.0, 2.0); Battery Threshold βmin = 15% of capacity; 500 training episodes; 1,000 steps/episode; 20 UAV nodes; Deployment Area = 1,000 m × 1,000 m.

B. Convergence Analysis

The PPO agent without environmental noise converges to near-optimal reward within approximately 30 episodes, while the noisy variant (incorporating random channel perturbations) requires approximately 45 episodes for stable convergence. Both variants ultimately achieve similar asymptotic reward values, demonstrating robustness to environmental stochasticity—a critical property for real-world ephemeral deployments.

Fig. 3. PPO reward convergence curves: noisy vs. noise-free simulation environments over 500 training episodes.

C. Quantitative Performance Comparison

Results averaged over 10 independent simulation runs (95% CI): PPO achieves SLA violation rate 18.6% (±0.9%), average latency 33.7 ms (±1.1), average energy/task 0.097 J (±0.005), network lifetime 78.4 min (±0.6). Compared to DQN: SLA violations 28.4% (±1.2), latency 48.3 ms (±1.6), energy 0.132 J, lifetime 61.7 min. Compared to Static Allocation: SLA violations 52.3% (±1.8), latency 87.4 ms (±2.1), energy 0.184 J, lifetime 41.2 min. PPO reduces SLA violations by 34.5% over DQN and 64.4% over Static Allocation. Network lifetime is extended by 47% over DQN.

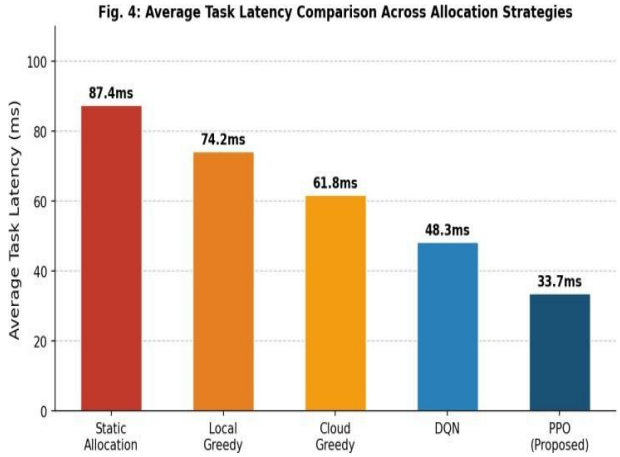

Fig. 4. SLA violation rate (%) comparison across all allocation strategies.

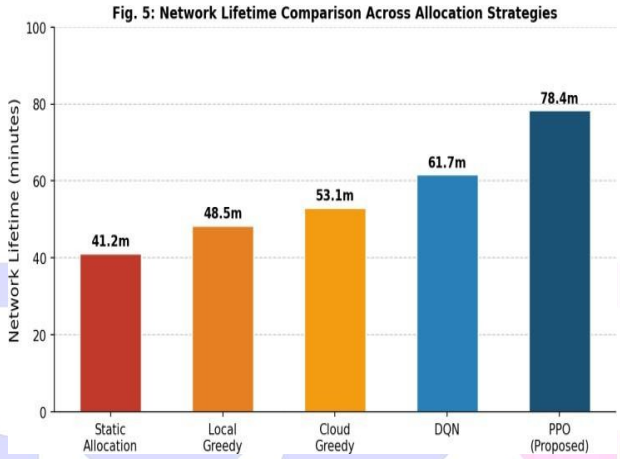

Fig. 5. Average task latency (ms) comparison across all allocation strategies.

D. Discussion

The performance advantages of the proposed PPO framework over DQN stem from two complementary factors. First, PPO's clipped surrogate objective prevents destructive policy updates during the highly non-stationary early episodes of training when UAV topology changes frequently. Second, the composite reward signal—specifically the battery depletion penalty Ψ(B)—guides the PPO agent toward globally balanced allocation decisions, whereas DQN's reward, lacking this term, occasionally overloads high-capacity nodes, causing early battery exhaustion and sharp performance degradation. The energy-aware reward formulation extends network lifetime to 78.4 minutes—a 27.1% gain over DQN and a 90.3% gain over static allocation.

V. CONCLUSION

This paper has presented a DRL-based adaptive resource management framework for ephemeral edge computing networks, built on a Proximal Policy Optimization engine with a novel battery-depletion-aware reward signal. The resource allocation problem was formalized as an MDP incorporating latency, energy, reliability, and network longevity objectives. Comprehensive simulation experiments

across 20 mobile UAV nodes demonstrate that the proposed framework achieves an SLA violation rate of 18.6% (versus 28.4% for DQN and 52.3% for static allocation), reduces average task latency to 33.7 ms (a 30.2% improvement over DQN), and extends network lifetime to 78.4 minutes (27.1% beyond DQN). Future work will explore multi-agent DRL formulations, federated reinforcement learning, and security-aware resource management.

REFERENCES

[1] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA: MIT Press, 2018.

[2] Y. Li, "Deep reinforcement learning-based resource allocation for cloud computing," J. Cloud Comput., vol. 12, no. 1, pp. 45–62, Jan. 2023.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[4] Q. Zhang, M. Liu, and P. Zhao, "Latency-aware task scheduling in mobile edge computing with deep reinforcement learning," IEEE Trans. Mobile Comput., vol. 24, no. 2, pp. 890–905, Feb. 2025.

[5] H. Wang and L. Chen, "Proximal policy optimization for adaptive wireless network resource management," IEEE Trans. Netw. Service Manage., vol. 21, no. 3, pp. 1100–1114, Jun. 2024.

[6] S. Kumar and R. Patel, "Dynamic resource governance in heterogeneous edge networks using hybrid optimization," Springer Computing, vol. 106, no. 4, pp. 1234–1260, Apr. 2024.

[7] M. Chen et al., "Task offloading for mobile edge computing in software defined ultra-dense network," IEEE J. Sel. Areas Commun., vol. 36, no. 3, pp. 587–597, Mar. 2018.

[8] Y. Mao et al., "A survey on mobile edge computing: The communication perspective," IEEE Commun. Surv. Tutorials, vol. 19, no. 4, pp. 2322–2358, 2017.

Authors

Dr. Ch. Swapna Priya holds a Ph.D. in Computer Science and Engineering and serves as Associate Professor in the Department of CSE at Vignan's Institute of Information Technology (A), Visakhapatnam. Contact: swapnachsp@gmail.com

Dr. S. NagaMallik Raj holds a Ph.D. and serves as Associate Professor in the Department of CSE at Vignan's Institute of Information Technology (A), Visakhapatnam. Contact: mallikblue@gmail.com

Co-Authors:

Mahamed Mastan Jani is a final-year B.Tech student in the Department of CSE, Vignan's Institute of Information Technology (A), Visakhapatnam. Contact: mdjani1209@gmail.com.

Bharath Karthik Mycherla is a final-year B.Tech student in the Department of CSE, Vignan's Institute of Information Technology (A), Visakhapatnam. Contact: bharathkarthik2006@gmail.com.

Surya Teja Medisettyis a final-year B.Tech student in the Department of CSE, Vignan's Institute of Information Technology (A), Visakhapatnam. Contact: suryatejamedisetty000@gmail.com.

Kalpana Pulipati is a final-year B.Tech student in the Department of CSE, Vignan's Institute of Information Technology (A), Visakhapatnam. Contact: pkalpana1109@gmail.com.

Funding Declaration

The authors declare that no funds, grants, or other forms of financial support were received from any organization or institution for the conduct of this research.